時間平均多重連鎖セミ・マルコフ決定過程における修正政策反復法の収束について(学習と制御とその周辺)

AUTHOR(S):

大野, 勝久

CITATION:

大野, 勝久. 時間平均多重連鎖セミ・マルコフ決定過程における修正政策反復法の収束について(学習と制御とその周辺). 数理解析研究所講究録 1985, 557: 79-94

ISSUE DATE:

1985-04

URL:

http://hdl.handle.net/2433/98984

数理解析研究所講究録
557巻 1985年 79-94

時間平均多重連鎖セミ・マルコフ決定過程における修正政策反復法の収束について

甲南大学　理学部　大野勝久(Katsuhisa Ohno)

1. はじめに

有限状態，有限決定セミ・マルコフ決定過程において時間平均利得を最大にする最適定常政策を決定するアルゴリズムとしては，政策反復法（PIM），逐次近似法（SAM），線形計画法（LP），修正政策反復法（MPIM）が知られている[1]。特にPIMはよく知られているが，PIM，LPは共に多状態問題にたいしては適用が困難であり，これら問題にたいしてはSAMおよびその一般化としてのMPIMが有力な手法として研究されてきた。しかしSAM，MPIMは多重連鎖問題にたいしては無力であった。最近 Schweitzer [2] は多重連鎖問題にたいするSAMを提案し，その収束を示している。しかしながらそのアルゴリズムは複雑であり，実用的とは思われない。本論文ではより一般的にMPIMの収束について論ずる。

2. 政策反復法

以下の記号を使用する。

$I = \{1, 2, \cdots, M\}$：状態空間

K_i $(i \in I)$：状態 i でとりうる決定の有限集合

$r_i(k)$ $(i \in I,\ k \in K_i)$：i において k をとって えられる平均利得

$p_{ij}(k)$ $(i, j \in I,\ k \in K_i)$：i において k をとったとき j へ遷移する確率

$T_i(k)$ $(i \in I,\ k \in K_i)$：平均遷移時間 $(T_i(k) > 0$ とする$)$

F：定常政策 $f = (f_1, f_2, \cdots, f_M)$ の集合 $(f_i \in K_i,\ i \in I)$

$r(f) = (r_1(f_1), r_2(f_2), \cdots, r_M(f_M))^T$ （T は転置を表わす）

$P(f) = (p_{ij}(f_i))$：f にたいする遷移確率行列

$T(f) = \mathrm{diag}\,(T_i(f_i))$

$g(f)$：定常政策 f のゲイン

$v(f)$：定常政策 f の相対値

$g(f)$, $v(f)$ は，$P(f)$ の各エルゴード部分連鎖に属する 1 つの状態 s にたいして $v_s = 0$ とおいて，次の連立一次方程式を解いて えられる。

$$g = P(f)\,g \quad , \quad v = r(f) + P(f)\,v - T(f)\,g \qquad (1)$$

ここで，$0 < \alpha < 1$，

$$0 < \tau \leq (1-\alpha) \min_{i,k} \{T_i(k)/(1-p_{ii}(k))\ ;\ p_{ii}(k) < 1\} \qquad (2)$$

をみたす任意の α, τ にたいして

$$\Omega(f) = \tau T(f)^{-1} \tag{3}$$

とおき，Schweitzer の data 変換：

$$\hat{P}(f) = I - \Omega(f) + \Omega(f)P(f), \quad \hat{r}(f) = \Omega(f)r(f) \tag{4}$$

を用いれば (1)式は

$$g = \hat{P}(f)g, \quad v = \hat{r}(f) + \hat{P}(f)v - g \tag{5}$$

となる。ただし，(4)式の I は単位行列であり，(5)式の解 $\hat{g}, \hat{v}$ は $\hat{g} = \tau g(f)$, $\hat{v} = v(f)$ となり，任意の $f \in F$ にたいして

$$\hat{P}(f) \geq \alpha I \tag{6}$$

をみたす。すなわち，セミ・マルコフ決定過程は上記 data 変換で任意の $f \in F$ にたいして (6)式をみたす同値なマルコフ決定過程に変換される。

以下記号を簡略化するため変換されたマルコフ決定過程にたいして再び $r(f)$, $P(f)$ を用いることにする。すなわち

$$g = P(f)g, \quad v = r(f) + P(f)v - g \tag{7}$$

の解が $g(f)$, $v(f)$ であり，$P(f)$ は任意の $f \in F$ にたいして

$$P(f) \geq \alpha I \tag{8}$$

をみたす。また最大ゲイン $g^* = g(f^*)$, $v^* = v(f^*)$ のみたす最適方程式は

$$g_i^* = \max_{k \in K_i} \{ \sum_{j \in I} P_{ij}(k) g_j^* \} \quad (i \in I) \tag{9}$$

$$g_i^* + v_i^* = \max_{k \in L_i} \{ r_i(k) + \sum_{j \in I} p_{ij}(k) v_j^* \} \quad (i \in I) \qquad (10)$$

で与えられる。ここで $L_i = \{k \in K_i ;$ (9) 式右辺を最大化する $k\}$ である。

[補題1]

$P(f^*)$ のエルゴード部分連鎖の状態集合を E_r^* $(r=1, \cdots, R^*)$, 過渡状態の集合を T^* とし, $g_r^* = g_i(f^*)$ $(i \in E_r^*)$ とおく。

$$g_1^* < g_2^* < \cdots < g_R^*$$

ならば,

i) $T^* \ni i$ について $g_1^* \leq g_i(f^*) \leq g_R^*$ である。

ii) $E_r^* \ni i$ について

$$f_i \in D_i = K_i - \{k \in K_i ; E_1^* \cup \cdots \cup E_{r-1}^* \cup \{i \in T^* ; g_i(f^*) \leq g_r^*\} \ni j \text{ について } p_{ij}(k) > 0\}$$

であり, E_r^* は任意の $f_i \in D_i$ $(i \in E_r^*)$ で閉じている。

(証明) i) は明らかであり, $f_i \notin D_i$ $(i \in E_r^*)$ ならば E_r^* がエルゴード部分連鎖であることと矛盾することから ii) の前半が示される。E_r^* が閉じていない $f_i \in D_i$ $(i \in E_r^*)$ が存在したとすれば, (9) 式より

$$g_r^* = \max_{k \in K_i} \{ \sum_{j \in I} p_{ij}(k) g_j(f^*) \} \geq g_r^* \sum_{j \in E_r^*} p_{ij}(f_i) + \sum_{j \notin E_r^*} p_{ij}(f_i) g_j(f^*) > g_r^*$$

となり, 矛盾である。

多重連鎖マルコフ決定過程の最適定常政策 f^* を求める最もよく知られた手法は次の Howard の政策反復法である。

1. 初期政策 f^0 を与え，$n=0$ とおく．

2.（値決定ルーチン）

$P(f^n)$ の各エルゴード部分連鎖に属する 1 つの状態 s で $v_s=0$ とおき，(7) 式を解いて $g(f^n)$，$v(f^n)$ を定める．

3.（政策改良ルーチン）各 $i\in I$ で

$L_i^{n+1}=\{\sum_{j\in I}p_{ij}(k)g_j(f^n)$ を最大化する $k\in K_i\}$

$F_i^{n+1}=\{r_i(k)+\sum_{j\in I}p_{ij}(k)v_j(f^n)$ を最大化する $k\in L_i^{n+1}\}$

を定め，$f_i^n\in F_i^{n+1}$ ならば $f_i^{n+1}=f_i^n$ とおき，さもなければ f_i^{n+1} を F_i^{n+1} の適当な要素ととる．全ての $i\in I$ で $f_i^{n+1}=f_i^n$ となれば停止．f^{n+1} は最適政策 f^* であり，最大ゲイン $g^*=g(f^n)$，相対値 $v^*=v(f^n)$ である．さもなければ $n=n+1$ とおいてステップ 2 へ．

3. 修正政策反復法

PIM の値決定ルーチンを有限回の逐次近似法でおきかえた手法が MPIM である．$F\ni f$ にたいして初期ベクトル w^0 ではじまる次の逐次近似法 $\{w^l;\ l=0,1,\cdots\}$ を考える．

$$w^{l+1}=r(f)+p(f)w^l. \tag{11}$$

このとき，$l=0,1,\cdots$ にたいして

$$w^l=lg(f)+v(f)+p(f)^l(w^0-v(f)) \tag{12}$$

がなりたち [3]，次の補題が成立する．

[補題 2]

$P(f)$ のエルゴード部分連鎖の状態集合を E_r $(r=1,\cdots,R)$，過渡状態の集合を T とし，$g_r = g_i(f)$ $(i \in E_r)$，$g_1 < g_2 < \cdots < g_R$ とする。このとき

i) $$\lim_{l\to\infty} w^{l+1} - w^l = g(f) \tag{13}$$

ii) $E_r \ni i, s_r$ について，$v_{s_r}(f) = 0$ であれば，

$$\lim_{l\to\infty} w_i^l - w_{s_r}^l = v_i(f) \tag{14}$$

であり，$s_p \in E_p$ $(p<r)$ であれば，

$$\lim_{l\to\infty} w_i^l - w_{s_p}^l = \infty . \tag{15}$$

iii) $$P(f) = \begin{pmatrix} Q(f) & 0 \\ R(f) & S(f) \end{pmatrix} \tag{16}$$

とする。ここで $Q(f)$ は $E_1 \cup \cdots \cup E_R$ に対応し，$R(f)$, $S(f)$ は T に対応する。このとき

$$v_i^l = w_i^l - w_{s_r}^l \quad (i \in E_r,\ r=1,\cdots,R) \tag{17}$$

$$g_i^l = w_i^l - w_i^{l-1} \quad (i \in I) \tag{18}$$

とおき，r_t, g_t で T に対応する部分ベクトルを表わし，u_i^{l+1} $(i \in T)$ を $u_i^1 = w_i^1$ $(i \in T)$ から

$$u^{l+1} = r_t(f) + R(f)v^l - g_t^l + S(f)u^l \tag{19}$$

で求めれば，$T \ni i$ について

$$\lim_{l\to\infty} u_i^l = v_i(f) . \tag{20}$$

がなりたつ。

（証明）（8）式より

$$\lim_{l\to\infty} P(f)^l = P^*(f) \equiv \lim_{N\to\infty} \frac{1}{N}\sum_{l=1}^{N} P(f)^l \quad , \quad P^*(f)P(f) = P(f)P^*(f) = P^*(f) \quad (21)$$

であるから，i）は（12）式より明らかである。ii）より E_r のえにたいして

$$g_r = \lim_{l\to\infty} w_i^{l+1} - w_i^l = \lim_{l\to\infty} r_i(f_i) + \sum_{j\in I} p_{ij}(f_i)(w_j^l - w_{sr}^l) - (w_i^l - w_{sr}^l) \quad (22)$$

であり，（7）式がなりたつことから（14）式がえられる。

（15）式は（12）式から

$$\lim_{l\to\infty} w_i^l - w_{sp}^l = l(g_r - g_p) + v_i(f) + p(f)^l(w^0 - v(f))_i - p(f)^l(w^0 - v(f))_{sp}$$

となることから導かれる。iii）の証明同様，（17），（18）式の v_i^l，g_i^l は $l\to\infty$ で $v_i(f)$ $(i\in E_r)$，$g_i(f)$ へ収束し，$S(f)^l \to 0$ $(l\to\infty)$ から（20）式がえられる。

PIMの値決定ルーチンに上記逐次近似法を用いればMPIMがえられる。しかし補題1で述べた一般的な多重連鎖問題にたいするMPIMは複雑であり，以下では簡単のため $R^* \leq 2$，すなわち $P(f^*)$ がたかだか2つのエルゴード部分連鎖をもつことが知られた問題にたいするMPIMを述べ，次節でその収束を示す。

修正政策反復法（$R^* \leq 2$）

1．初期ベクトル v^0，非負整数 m，大きな数 L，小さな正数 ε, δ を定め，$n = c = 0$，$E_1 = I$，$T = \phi$，$R^* = 1$，$D_i = K_i$ $(i\in I)$ とおく。

2．（政策改良ルーチン）

i) $i \in E_r$ $(1 \le r \le R^*)$ にたいして

$$x_i^{n+1} = \max_{k \in D_i} \{ r_i(k) + \sum_{j \in I} p_{ij}(k) v_j^n - v_i^n \} \tag{23}$$

を計算し，f_i^n が最大値を与えれば $f_i^{n+1} = f_i^n$ とおき，さもなければ f_i^{n+1} を最大値を与える任意の $k \in D_i$ とする。

$s_r = \arg \min_{i \in E_r} \{ x_i^{n+1} \}$ とおく。

ii) $i \in T$ にたいして

$$L_i^{n+1} = \{ k \in K_i ; \sum_{j \in I} p_{ij}(k) g_j^m \ge \max_{k \in K_i} \{ \sum_{j \in I} p_{ij}(k) g_j^m \} - \delta \} \tag{24}$$

$$x_i^{n+1} = \max_{k \in L_i^{n+1}} \{ r_i(k) + \sum_{j \in I} p_{ij}(k) v_j^n - v_i^n \} \tag{25}$$

を計算し，f_i^n が最大値を与えれば $f_i^{n+1} = f_i^n$ とおき，さもなければ f_i^{n+1} を最大値を与える任意の $k \in L_i^{n+1}$ とする。

3．（値決定ルーチン）

$$w^0 = v^n + x^{n+1} \tag{26}$$

とおき，$l = 0, 1, \cdots, m-1$ にたいして

$$w^{l+1} = r(f^{n+1}) + p(f^{n+1}) w^l \tag{27}$$

を計算し，

$$v_i^{n+1} = w_i^m - w_{s_r}^m \quad (i \in E_r,\ 1 \le r \le R^*) \tag{28}$$

$$g_i^m = w_i^m - w_i^{m-1} \quad (i \in I) \tag{29}$$

とおく。$u_i^0 = w_i^0$ $(i \in T)$ とおき，$l = 0, \cdots, m-1$ にたいし

$$u^{l+1} = r_t(f^{n+1}) + R(f^{n+1}) v^{n+1} - g_t^m + S(f^{n+1}) u^l \tag{30}$$

を計算し，v_i^{n+1} $(i \in T)$ を次式で定める。

$$v_i^{n+1} = u_i^m \tag{31}$$

4. $n=n+1$ とおき，$c=1$ ならばステップ2へ．$c=0$ ならば $\|v^n\|_d \equiv \max_{i\in I} v_i^n - \min_{i\in I} v_i^n$ をもとめ，$\|v^n\|_d < L$ ならばステップ2へ．さもなければステップ5へ．

5.（状態分類ルーチン）

s_1 を含み全ての $f_i \in K_i$ で閉じている状態の集合 E_1' および $\arg\max_{i\in I} v_i^n$ を含み f^n で閉じている状態の集合 E_2' を定める．$\max_{i\in E_2'} x_i^n - \min_{i\in E_2'} x_i^n < \varepsilon$ ならば $R^*=2$，$E_r = E_r'$ $(r=1,2)$，$T = I - E_1 - E_2$，$D_i = K_i - \{k\in K_i ;\ j\notin E_2 \text{ にたいして } p_{ij}(k)>0\}$ $(i\in E_2)$，$c=1$ とおき，ステップ2へ．さもなければ，$L=2L$ とおいてステップ2へ．

4．修正政策反復法の収束

まず $c=0$，すなわち $E_1 = I$，$T=\phi$，$D_i = K_i$ $(i\in I)$ にたいする収束を論ずる．(7)式より，(23)，(28)式は各々

$$x^{n+1} = g(f^{n+1}) - (I - P(f^{n+1}))(v^n - v(f^{n+1})) \tag{32}$$

$$v^{n+1} = (m+1)g(f^{n+1}) + v(f^{n+1}) + P(f^{n+1})^{m+1}(v^n - v(f^{n+1})) - w_{s_1}^m e \tag{33}$$

と書くことができる．ここで $e=(1,\cdots,1)^T$ である．

[補題3]

i)
$$x^{n+1} = d^{n+1} + P(f^n)^{m+1} x^n \tag{34}$$

$$d^{n+1} = \max_{f\in F}\{(m+1)[P(f)g(f^n) - g(f^n)] + r(f) + P(f)v(f^n) - g(f^n) - v(f^n) + (P(f) - P(f^n))P(f^n)^{m+1}(v^{n-1} - v(f^n))\} \geq 0 \tag{35}$$

$d_i^{n+1} = 0$ となる必要十分条件は $f_i^{n+1} = f_i^n$ である．

ii) $x^{n+1} \geqq p(f^n)^{m+1} x^n$, $x^{n+1} + v^n \geqq r(f^*) + p(f^*) v^n$ (36)

$p^*(f^*) x^{n+1} \geqq g(f^*) \geqq g(f^{n+1}) = p^*(f^{n+1}) x^{n+1}$, $p^*(f^n) x^{n+1} \geqq g(f^n)$ (37)

iii) $\Delta_n = \max_{i \in I} x_i^n$, $\nabla_n = \min_{i \in I} x_i^n$ (38)

とおけば

$$\Delta_n e \geqq g(f^*) \geqq g(f^{n+1}) \geqq \nabla_{n+1} e \geqq \nabla_n e \quad (39)$$

iv) 状態の集合 C が $p(f)^{m+1}$ で閉じていれば, $p(f)$ でも閉じている.

(証明) i) (12) 式と同様

$$v^n = m g(f^n) + r(f^n) + p(f^n)^m (v^{n-1} + x^n - r(f^n)) - w_{s,i}^m e$$

がなりたち, この式と (33) 式を (23) 式に代入し, (7), (32) 式を用いて整理すれば i) が示される. ii) は i), (21) 式より導かれ, iii), iv) は明らかである.

[補題4]

i) $\bar{\nabla} = \lim_{n \to \infty} \nabla_n$ が存在し, 空でない集合 $C \subset I$ にたいして

$$\lim_{n \to \infty} x_i^n = \bar{\nabla} \quad (i \in C). \quad (40)$$

ii) 有限な N が存在し, 全ての $n \geqq N$ にたいして

$$f_i^n = f_i \in K_i , \quad d_i^n = 0 \quad (i \in C) \quad (41)$$

であり, C は $p(f^n)$ で閉じている.

(証明) (39) 式より ∇_n は $\bar{\nabla}$ へ収束する. $i \in I$ にたいして, $y_i = \liminf_{n \to \infty} x_i^n$, $z_i = \limsup_{n \to \infty} x_i^n$ とおき,

$A = \{ i \in I ;$ 無限に多くの n で $x_i^n = \nabla_n \}$

とおけば A は空でなく，$A \ni i$ にたいして $y_i = \bar{\nabla}$ がなりたつ．

したがって

$$C = \{ i \in I \ ; \ y_i = \bar{\nabla} \}$$

とおけば C は空でない．定義より $i \in C$ にたいして $\{n\}$ の部分列 $L(i) = \{l\}$，$U(i) = \{u\}$ が存在し，

$$\lim_{l \to \infty} x_i^l = \bar{\nabla} \quad , \quad \lim_{u \to \infty} x_i^u = z_i$$

である．部分列 $\{ n_k \ ; \ k = 0, 1, \cdots \}$ を

$$n_{2k} \in U(i), \ n_{2k+1} \in L(i), \ n_{2k-1} < n_{2k} < n_{2k+1}, \ n_{2k+1} - n_{2k} < \infty$$

で構成し，$P(2k) = P(f^{n_{2k+1}-1})^{m+1} \cdots P(f^{n_{2k}})^{m+1}$ とおけば，(36) 式より

$$x^{n_{2k+1}} \geqq P(f^{n_{2k+1}-1})^{m+1} x^{n_{2k+1}-1} \geqq \cdots \geqq P(2k) x^{n_{2k}}$$

である．ゆえに

$$x_i^{n_{2k+1}} \geqq P_{ii}(2k) x_i^{n_{2k}} + \sum_{j \neq i} P_{ij}(2k) x_j^{n_{2k}} \geqq \nabla_{n_{2k}} + P_{ii}(2k)(x_i^{n_{2k}} - \nabla_{n_{2k}})$$

がなりたち，(8) 式より $P_{ii}(2k) > 0$ であるから $z_i \leq \bar{\nabla}$ となり，i) が示される．

ii) $C \ni i$ にたいして十分大きな n で

$$x_i^n > \bar{\nabla} + \gamma \tag{42}$$

となる $\gamma > 0$ が存在する．(36) 式より $i \in C$ にたいして

$$x_i^{n+1} \geqq \nabla_n + \sum_{j \notin C} P_{ij}(f^n)^{m+1}(x_j^n - \nabla_n) \geqq \nabla_n + \gamma \sum_{j \notin C} P_{ij}(f^n)^{m+1}$$

である．ゆえに i) から $\sum_{j \notin C} P_{ij}(f^n)^{m+1} = 0$ であり，補題 3-iv) より C は $P(f^n)$ で閉じている．したがって (34) 式より

$$\lim_{n\to\infty} d_i^{n+1} = \lim_{n\to\infty} \left(x_i^{n+1} - \sum_{j\in C} P_{ij}(f^n)^{n+1} x_j^n \right) = 0$$

となり，補題 3－i)から f_i^n はある決定 $\bar{f}_i \in K_i$ へ収束する．C，K_i ともに有限であるから ii) がなりたつ．

[定理 1]

$C = I$ ならば，I のえにたいして

$$f_i^n = f_i^* \quad (n \geq N), \quad \lim_{n\to\infty} x_i^n = g_i^* \quad , \quad \lim_{n\to\infty} v_i^n = v_i(f^*)$$

がなりたつ．

(証明) 補題 3－ii) および 4 よりただちに

$$g(f^*) = g_i^* e = g(\bar{f}) = \bar{v} e \tag{43}$$

がえられる．任意の $\varepsilon > 0$ にたいして $\bar{N}(\geq N)$ が存在し，$n \geq \bar{N}$ において

$$x^n = \bar{v} e + \varepsilon^n \quad , \quad |\varepsilon_i^n| < \varepsilon \tag{44}$$

であるから

$$g(\bar{f}) + v^n + \varepsilon^n = r(\bar{f}) + p(\bar{f}) v^n$$

であり，(7)式より

$$\lim_{n\to\infty} v^n = v(\bar{f}) \tag{45}$$

がなりたつ．ゆえに $d^n = 0$ より

$$g(\bar{f}) + v(\bar{f}) = \max_{f\in F} \{ r(f) + p(f) v(\bar{f}) \}$$

となり，$\bar{f} = f^*$ が示される．

以下 $C \neq I$ の場合を考える．$T = I - C$ とおけば補題 4 より $n \geq N$ において $p(f^n)$ は

$$P(f^n) = \begin{pmatrix} Q(\bar{f}) & 0 \\ R(f^n) & S(f^n) \end{pmatrix} \tag{46}$$

と表わすことができる。ここで R, S は T に対応する行列である。以下添字 c, t で C, T に対応する部分ベクトルを表わすことにする。

[定理2]

$C \neq I$ ならば $K^* = 2$, $g_1^* < g_2^*$, $C = E_1^*$ であり, E_3^* のまわりにたいして $f_i^n = f_i^*$ $(n \geq N)$, $\lim_{n\to\infty} x_i^n = g_1^*$, $\lim_{n\to\infty} v_i^n = v_i(f^*)$ がなりたつ。

(証明) 補題3, 4 より

$$g_c(f^n) = Q^*(\bar{f})x_c^n = Q^*(\bar{f})v_c(\bar{f}) = g_c(\bar{f}) = \bar{\nu} e_c \tag{47}$$

がなりたち, $n \geq \bar{N}$ で

$$x_c^n = \bar{\nu} e_c + \varepsilon_c^n, \quad |\varepsilon_i^n| < \varepsilon \quad (i \in C) \tag{48}$$

である。ゆえに定理1の証明と同様にして (45) 式がなりたち,

$$v_c^n = v_c(\bar{f}) + \delta_c^n = v_c(f^{n+1}) + \delta_c^n, \quad |\delta_i^n| < \varepsilon \quad (i \in C) \tag{49}$$

である。(32) 式から

$$x_t^{n+1} = g_t(f^{n+1}) - R(f^{n+1})(v_c^n - v_c(f^{n+1})) - (I_t - S(f^{n+1}))(v_t^n - v_t(f^{n+1}))$$

であるから (49) 式より

$$x_t^{n+1} - g_t(f^{n+1}) + R(f^{n+1})\delta_c^n = (I_t - S(f^{n+1}))(v_t(f^{n+1}) - v_t^n) \tag{50}$$

となり,

$$(I_t - S(f^n)^{m+1})(v_t(f^{m+1}) - v_t^n) = \sum_{\ell=0}^{m} S(f^{m+1})^{\ell}\{x_t^{m+1} - g_t(f^{m+1}) + R(f^{m+1})\delta_c^n\} \qquad (51)$$

がなりたつ。一方，(33) 式に (46), (47), (49) 式を代入すれば，

$$v_c^{n+1} = (m+1)\bar{\nabla}e_c + v_c(\bar{f}) + Q(\bar{f})^{m+1}\delta_c^n - w_{S_1}^m e_c \qquad (52)$$

$$v_t^{n+1} = (m+1)g_t(f^{m+1}) + v_t(f^{m+1}) + R(f^{m+1})^{(m+1)}\delta_c^n + S(f^{m+1})^{m+1}(v_t^n - v_t(f^{m+1})) - w_{S_1}^m e_t \qquad (53)$$

となる。ここで $R(f)^{(m+1)} = \sum_{\ell=0}^{m} S(f)^{\ell} R(f) Q(\bar{f})^{m-\ell}$ である。$v_{S_1}^{n+1} = v_{S_1}(\bar{f}) = 0$ であるから (52) 式より $w_{S_1}^m = (m+1)\bar{\nabla} + (Q(\bar{f})^{m+1}\delta_c^n)_{S_1}$ であり，(53) 式は

$$v_t^{n+1} - v_t(f^{m+1}) = (m+1)(g_t(f^{m+1}) - \bar{\nabla}e_t) + R(f^{m+1})^{(m+1)}\delta_c^n - (Q(\bar{f})^{m+1}\delta_c^n)_{S_1}e_t + S(f^{m+1})^{m+1}(v_t^n - v_t(f^{m+1}))$$

となる。したがって，(42), (51) 式を用いて整理すれば，

$$v_t^{n+1} - v_t^n = \sum_{\ell=0}^{m} S(f^{m+1})^{\ell}\{x_t^{m+1} - \bar{\nabla}e_t + R(f^{m+1})(I_c + Q(\bar{f})^{m-\ell})\delta_c^n\} - (Q(\bar{f})^{m+1}\delta_c^n)_{S_1}e_t$$

$$> (\gamma - 3\varepsilon)e_t + \gamma\sum_{\ell=1}^{m} S(f^{m+1})^{\ell}e_t + 2\varepsilon S(f^{m+1})^{m+1}e_t \qquad (54)$$

がなりたち，$\gamma > 3\varepsilon$ とすれば

$$\lim_{n\to\infty} v_i^n = \infty \quad (i \in T) \qquad (55)$$

がえられる。ある $i \in C$ で $P_{ij}(f_i^*) > 0$ $(j \in T)$ と仮定すれば，(36) 式 2 式右辺が ∞ となり矛盾である。すなわち任意の $f \in F$ で C は閉じており，

$$g_c(\bar{f}) = \max_{f\in F}\{P_c(f)g_c(\bar{f})\},\ g_c(\bar{f}) + v_c(\bar{f}) = \max_{f\in F}\{r_c(f) + P_c(f)v_c(\bar{f})\}$$

がなりたつ。ゆえに $R^*=2$, $g_1^*<g_2^*$, $C=E_1^*$, $\bar{f}_i=f_i^*$ $(i\in E_1^*)$ が示される。

定理1, 2より $R^*\leq 2$ にたいするMPIMの収束をある条件のもとで示すことができる。

[定理3]

修正政策及後法は, $R^*=1$ あるいは $R^*=2$, $g_1^*=g_2^*$ ならば常に, $R^*=2$, $g_1^*<g_2^*$ ならば,

i) $\bar{N}$ が存在し, E_2^* は $n\geq\bar{N}$ となる全ての n において $P(f^n)$ で閉じている。

ii) T^* の i にたいして, $\sum_{j\in T^*} P_{ij}(f_i^*)<1$ がなりたつ。

の条件のもとで、任意の v^0, m にたいして収束する。

(証明) 定理1, 2より $R^*=1$ あるいは $R^*=2$, $g_1^*=g_2^*$ ならば $C=I$ となり, MPIMは任意の v^0, m で収束する。$R^*=2$, $g_1^*<g_2^*$ ならば定理2より $C=E_1^*$ であり, 仮定i)から $n\geq\bar{N}$ において $f_i^n=f_i^*$ $(i\in E_1^*)$, $\sum_{j\in E_2^*} P_{ij}(f_i^n)=1$ $(i\in E_2^*)$ がなりたつ。さらに E_1^* の i にたいして

$$\lim_{n\to\infty} x_i^n=g_1^* \quad , \quad \lim_{n\to\infty} v_i^n=v_i(f^*)$$

がなりたつ。仮定i)より $n\geq\bar{N}$ において x_i^n $(i\in E_2^*)$ は K_i を $D_i=K_i-\{k\in K_i;\ j\notin E_2^*$ にたいして $P_{ij}(k)>0\}$ に限定してえられるものと一致する。したがって定理1で, $I=E_2^*$ とおけば,

$$\lim_{n\to\infty} x_i^n = g_2^* \qquad (i \in E_2^*)$$

がなりたつ。$\|v^n\|_d \uparrow \infty$, $\max_{i\in E_2} x_i^n - \min_{i\in E_2} x_i^n \to 0$ $(n\to\infty)$ であるから，ある有限な $n = N'$ でステップ5の条件が成立し，$E_1 = E_1^*$, E_2 が定められる。さて，仮定 i) より $E_2 \subset E_2^*$ であり，(37) 式より $i \in E_2$ にたいして

$$\varepsilon > \max_{i\in E_2} x_i^n - \min_{i\in E_2} x_i^n \geq \sum_{j\in E_2} p_{ij}^*(f^*)\{(\max_{i\in E_2} x_i^n) - x_j^n\}$$

であるから，$\sum_{j\in E_2} p_{ij}^*(f^*) = 0$ となり $E_2 \supset E_2^*$ がなりたつ。ゆえに $E_2 = E_2^*$ であり，$i \in E_2^*$ にたいする収束は定理1より導かれる。$i \in T = T^*$ にたいする収束は，仮定 ii) より K_i を $\sum_{j\in T^*} p_{ij}(k) < 1$ となる k に限定でき，割引利得問題にたいする収束証明 [4] と同様にして示すことができる。

参考文献

[1] 大野，マルコフ決定過程の計算アルゴリズム，第4回数理計画シンポジウム論文集 pp. 143-158, 1983.

2 Schweitzer,P.J., "A value-iteration scheme for undiscounted multichain Markov renewal programs," Zeitschrift für Operations Res. 28, 143-152, 1984.

3 Denardo,E.V., "Computing a bias-optimal policy in a discrete-time Markov decision problem," Opns. Res. 18, 279-289, 1970.

4 Ohno,K., "A unified approach to algorithms with a suboptimality test in discounted semi-Markov decision processes,"JORSJ 24, 296-324, 1981.